

# Birds of a Feather Session on Education and Curriculum for Data Curation

Attendees: Margaret Hedstrom, Zaman Aktaruzzaman, Stacy Kowalczyk, G. Sayeed Choudhury, Stacy Konkiel, Sharon Traweek

- MH gave background on her research project
  - NRC-backed research trying to get at the root of what skills are needed by "data curators", according to employers. Also seeking to find out about anticipated growth of field, experiences with difficulty/ease of hiring and training. Several meetings will happen throughout 2012, with a final report expected by the end of 2012.
  - They define data curation as "active management and enhancement of digital information for current and future use"
  - Report will produce both research questions and a set of recommendations to existing employers. Will help employers and educators articulate what we want the field to look like in the future, what kind of workforce is needed, and how big the field is going to be.
- Some questions to get us started:
  - What basic knowledge and skills are needed by data curators (whether in private or public sector)?
  - How much knowledge used by data curators (DC) is domain specific? What is the nature of that knowledge/skillset?
    - Sayeed pointed out that often DCs are appointed by research teams, not necessarily because they possess a certain skillset.
    - MH asked what comes first for DCs/information professionals who work with data? Are they subject specialists who go on to get a library science/information science degree? Or are they librarians/info science workers who learn about the field and then dive into the data?
      - If the latter, how do you learn to be conversant in the subject you're curating data for?
- SKowal: Curation styles/choices depend upon the nature of the science/research. With "big science," it is easier to insert DC standards at the beginning of the research life cycle
- How is data curated, exactly?
  - SKowal: Can a large group of data from varying fields be considered a "collection," just because it is all data? Points out that IU, there's much opposition to that.
  - MH agrees that data can't be a "collection" unto itself because of the variety of sources.
  - SKowal points out that data isn't necessarily "curated" in the traditional sense; ad hoc data collection/contributions to repositories are influencing data curation more so than the way in which traditional collection specialists collect resources

- Sayeed thinks of data as being similar to an archival collection: fragile, primary source, tells a lot about what's being studied at the time
- Data life cycle depends upon a research project's life cycle - scientists aren't as protective of data when they are nearing the end of their project or grant. We need to embed the idea of curation "upstream," at the beginning of a project for this reason, so that sharing is easier when they do decide to do so.
- The recent Presidential State of the Union Address mentions "data management" as an important area of growth.
- To what extent are there overlaps in how data is managed between academia and the commercial sector? -- There are a lot of differences in norms and regulations between the two areas.
  - Commercial sector isn't thinking of preservation so much as storage and access
  - Sayeed shared an anecdote from a colleague in the commercial sector, who said that there are data storage concerns (storage is cheap, management is not), workforce issues (not enough good applicants), and difficulty of preserving variety of formats
- Z: What are the best approaches and policies to preserve large volumes of data for young researchers?
  - If you have to deposit your dissertation electronically, ask if also deposit your data. Ask about data collections/repositories in your subject of study.
  - There is a negative correlation between age and willingness to participate in data sharing
    - Therefore, DC are trying to work with younger researchers for early "upstream" data curation
  - JHU gets researchers to sign institutional agreements about data management when they help researchers with their funder data management plans, recommend that researchers deposit early and often (embargoes are possible for researchers who don't want to share early data)
  - JHU doesn't characterize data services as "library services" but "research services"
    - They embed training/education on data management into the existing research admin processes for the university in order to get largest base of educated researchers; they have partners in their research administration office who help publicize events related to data curation
  - New push from NSF for doctoral students: research integrity training, and taking care of data is part of that training
- Variety of academic fields are getting funding from NSF to design new subfields (mostly Masters programs), such as "bioinformatics, astroinformatics, etc"
  - There is a business opportunity in that Masters Students generally pay the full amount for study (rather than PhDs)
- Information studies programs are also seeing an opportunity to expand their services to allow study in these fields

- Trust
  - Trust vs subject-specific skills: which is more important?
  - A good technologists skills can help build trust. A person with lots of domain knowledge (and therefore has trust in field) doesn't need to know technology as deeply.
  - ST: Astronomers/physicists don't want outside information professionals to do DC, they want people with training in their own field to handle their data. Many believe that the knowledge transfer (for the subject specialty) is too difficult
  - SKowal: Anecdote about a colleague who believed that skills necessary for his workforce were not technical but attitudinal (creativity, lifetime learner, etc)
  - It is a challenge for library schools to attract folks with subject-specific skills in the sciences
- Traweek study - women/minorities/immigrants and their roles as data curators
  - In the physical sciences, she found that most of those "moved to the side" in order to take care of data are women, minorities, and immigrants
  - Evidence that this has been an opportunity for women to get a job in the field where positions of significant authority and standing have only been available to women within the last generation. Data curation position offers stability (you don't have to move around from institution to institution in order to advance), which can fit in with career/family puzzle that a lot of women struggle with
  - Is "moved to the side" for data curation an opportunity or a dead end, in terms of one's career?
    - Data management/curation is becoming an increasingly crucial role on research team staffs, but DC has been considered "crucial but secondary"-women and minorities gravitate (for whatever reason) towards those roles
  - Another difficult question: to what degree is DC a professional, unified career and to what degree as it grows and becomes more common is it a package of skills that becomes sub-divided into technical, professional, outsourced roles (and accompanying positions)? To what extent can DC be automated and to what extent does it need to be manual?
- Takeaways:
  - Lots of different models as to whether DC is embedded in research or if its a service that someone else provides
  - There are implications for what kind of people take on curation responsibilities, how they learn what they learn. "Insiders" to fields vs "Outsiders" (whether that is judged by one's grounding in the domain or their status as women, minorities, and immigrants)
  - There is a potential branch off of DC/data management into its own career path, divorced from subject specialties